BMI 713 / GEN 212

# Lecture 5: Inference on Proportions

- Sampling distribution of proportions
- Confidence intervals
- Hypothesis testing

October 7, 2010

## Recall: Binomial Distribution

- Binomial distribution
  - two categories: "success" and "failure"
  - each trial is independent with probability *p*
  - a fixed number of trial
- If a random experiment has two possible outcomes and we do *n* independent repetitions with identical success probability *p*, then *X ~ Bin(n,p)* and

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- The probability of obtaining at least *k* successes is

$$P(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

## Bionomial Distribution

- "Exact" methods calculate sum of the discrete probabilities to compute p-values
- This is not feasible or necessary for large sample sizes
- In such cases, we use the normal approximation to the bionomial distribution

- What are the mean and variance of the normal approximation?
- ***E(X) = np***
- ***Var(X) = npq*** where ***q=1-p***

## Estimation of a Population Proportion

- Example: We would like to estimate *p*, the probability that a person under the age of 40 who is diagnosed with lung cancer survives for at least 5 years
- A random sample of *n = 70* individuals is selected from the population
- It is found that only *X = 8* patients out of the 70 survive for 5 years
- The number of "successes" X has a **binomial distribution**
- Hypothesis testing can be done using the "exact" method for small sample sizes.

## Sampling Distribution of Proportions

- The 5-year survival probability is estimated by the sample proportion of individuals who survive for 5 years

$$\hat{p} = \frac{X}{n} = \frac{8}{70} = 0.114$$

- If repeated samples of size 70 are selected from the population, what is the **sampling distribution of proportions**?
- The mean: $E(\hat{p}) = p$
- The variance: $Var(\hat{p}) = \frac{pq}{n}$

## Normal Approximation

- We apply the central limit theorem to the binomial distribution.
- If we draw samples of size *n* from a population whose proportion of interest is *p*, then the sample proportions p will be approximately distributed as

$$\hat{p} \sim N(p, pq/n)$$

provided that *n* is large enough, e.g., $npq \geq 5$

Since p and q are not know, replace them with $\hat{p}$ and $\hat{q}$

## Back to the Example

- For the lung cancer 5-year survival data

$$n\hat{p}\hat{q} = 70(0.114)(0.886) = 7.1$$

- Since this is large enough, the distribution of $\hat{p}$ can be assumed to be normal:

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

is distributed approximately as N(0,1)

## Confidence Interval

- 95% confidence interval:

$$P(-1.96 \leq z \leq 1.96) = 0.95$$

- We substitute *z*

$$P\left(-1.96 \leq \frac{\hat{p} - p}{\sqrt{pq/n}} \leq 1.96\right) = 0.95$$

- Isolating *p* in the center

$$P\left(\hat{p} - 1.96\sqrt{\frac{pq}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{pq}{n}}\right) = 0.95$$

- Therefore, the 95% confidence interval is

$$\left(\hat{p} - 1.96\sqrt{\frac{pq}{n}}, \hat{p} + 1.96\sqrt{\frac{pq}{n}}\right)$$

## Confidence Interval

- Since $p$ is not known, we estimate this with

$$\left(\hat{p}-1.96\sqrt{\frac{\hat{p}\hat{q}}{n}},\hat{p}+1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

- For the proportion of individuals under the age of 40 who survive at least 5 years after being diagnosed with lung cancer is

$$\left(.114-1.96\sqrt{\frac{(.114)(.886)}{70}},.114+1.96\sqrt{\frac{(.114)(.886)}{70}}\right)$$

$$(0.041,0.188)$$

## Hypothesis Testing for One Proportion

- To test the null hypothesis

$$H_0 : p = p_0,$$

if $np_0q_0 \geq 5$ $(q_0 = 1 - p_0)$ then under $H_0$ the test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0q_0}{n}}}$$

is approximately normally distributed as N(0,1)

## Example

- Mendel: self-pollination of a pea plant that was heterozygous (Dd) for the dwarf gene would yield 3/4 tall plants and 1/4 dwarf plants. Among the F2 progeny from a cross of a tall (DD) and a dwarf (dd) plant, Mendel observed 787 tall plants and 277 dwarfs

H0: p=0.75 $\quad Z = \dfrac{787/1064 - .75}{\sqrt{\dfrac{.75(1-.75)}{1064}}} = -0.7788$

Thus, the p-value is 0.436 and we do not reject $H_0$

## Hypothesis Testing for Two Proportions

- Suppose two populations have unknown proportions $p_1$ and $p_2$ and we want to test

$$H_0 : p_1 = p_2$$
$$H_A : p_1 \neq p_2$$

- Take two samples of size $n_1$ and $n_2$, compute $\hat{p}_1$ and $\hat{p}_2$
- If $H_0$ is true, then both populations have the same proportion $p$, which we estimate as

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

## Proportion Test

- If $n_1\hat{p}_1(1-\hat{p}_1) \geq 5$ and $n_2\hat{p}_2(1-\hat{p}_2) \geq 5$

  then under $H_0$, the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  is distributed approximately as N(0,1)

## Example

- Pagano & Gauvreau Ch14
- An 18-month study on effectiveness of seat belts on mortality among pediatric victims of motor vehicle accidents.
- $H_0$: the proportions of children who die as a result of the accident are identical for the two groups wearing and not wearing seat belts
- In the sample of 123 wearing a seat belt, 3 died
- In the sample of 290 not wearing a seat belt, 13 died

## Example

$$\hat{p}_1 = x_1/n_1 = 3/123 = 0.024$$
$$\hat{p}_2 = x_2/n_2 = 13/290 = 0.045$$

- Is the discrepancy in sample proportions too large to be attributed to chance?

$$\hat{p} = (3+13)/(123+290) = 0.039$$
$$z = \frac{(0.024 - 0.045) - 0}{\sqrt{(0.039)(1-0.039)\left(\frac{1}{123} + \frac{1}{290}\right)}} = -1.01$$

- The p-value is 0.312. Therefore, we cannot reject the null hypothesis. This study does not provide evidence that the proportions of children dying differ between those who were wearing seat belts and those who were not.